Telecytology: Intraobserver and Interobserver Reproducibility in the Diagnosis of Cervical-Vaginal Smears

PATRICIA M. ALLI, MD, CURTIS W. OLLAYOS, MD,
LESTER D. THOMPSON, MD, IQBAL KAPADIA, MD,
DANIEL R. BUTLER, HT(ASCP), BRUCE H. WILLIAMS, DVM,
DOROTHY L. ROSENTHAL, MD, AND TIMOTHY J. O'LEARY, MD, PHD

Telecytologic diagnosis of cervical-vaginal smears is potentially useful because it could allow more efficient use of cytopathologist resources and expertise. A pathologist in one location could, in principle, review cytotechnologists' findings using a video display hundreds or thousands of miles away. Currently, bandwidth restrictions limit practical implementation of such a system to review of fields that had been selected for review by the cytotechnologist. The purpose of our investigation was to evaluate how well this type of review correlates with a review in which the entire slide is available for examination by the pathologist. We prospectively selected 100 consecutive cervical-vaginal smears over an 11-day period in August 1999. For each smear, 4 to 12 fields containing abnormal cells from each slide were digitally imaged. Each of 3 pathologists reviewed all digitized images and all glass slides. Diagnoses based on selected digitized images were compared with those based on conventional pathologist review. The κ statistic, a measure of chance-corrected agreement (reproducibility), was calculated in each setting. Overall, intraobserver and interobserver reproducibility of cervical-vaginal smear diagnoses is fair to excellent. The use of remote digital images for pathologist review did not introduce large (2-step) diagnostic disagreements. The disagreement between a pathologist's glass slide and digital diagnoses is less than that for different pathologists reviewing glass slides, although interobserver differences were even greater in the interpretation of digital images. HUM PATHOL 32: 1318-1322. This is a U.S. government work. There are no restrictions on its use.

Key words: **telecytology, diagnostic reproducibility, cervical-vaginal smears, Pap smear.**

Abbreviations: **RCC, reactive cellular changes; WNL, within normal limits; ASCUS, atypical squamous cells of undetermined significance; AGUS, atypical glandular cells of undetermined significance; LSIL, low-grade squamous intraepithelial lesion; HSIL, high-grade squamous intraepithelial lesion.**

Telepathology has been studied extensively as a means of diagnosis and consultation in surgical pathology.[1-11] In the discipline of cytology, telecytology has been proposed as a means of initial off-site diagnosis or expert consultation. Additionally, use of telecytology could be considered for the administration of proficiency testing. Previous authors[12] have studied the telecytologic diagnostic *accuracy* of cytotechnologists for cervical-vaginal smears and shown the accuracy of telecytology to be high (group crude agreement, 85.6%) but less than that of light microscopy (group crude agreement, 95.6%). However, before telecytology can be used confidently in the proposed scenarios, thorough evaluation of its true *diagnostic reproducibility* is needed. A recent study[13] examining intraobserver variability of 3 pathologists for the telecytologic diagnosis of breast fine-needle aspiration biopsy specimens found 80% to 96% diagnostic concordance between telecytologic diagnosis and glass slide diagnosis. The present study was undertaken to evaluate the diagnostic reproducibility of cervical-vaginal smears among 3 pathologists using digitally captured images and a light microscope.

MATERIALS AND METHODS

One hundred consecutive cervical-vaginal smears for which pathologist review was obtained were prospectively selected from the Armed Forces Institute of Pathology cytology service over an 11-day period in August 1999. Each cervical-vaginal smear consisted of 1 Papanicolaou-stained slide that had been screened by an experienced cytotechnologist. The cytotechnologist's diagnoses of the cases included 60 benign (37 reactive cellular changes [RCC] and 23 within normal limits [WNL]), 28 atypical squamous cells of undetermined significance (ASCUS), 1 atypical glandular cells of undetermined significance (AGUS), 9 low-grade squamous intraepithelial lesions (LSILs), and 2 high-grade squamous intraepithelial lesions (HSILs). The diagnoses were rendered utilizing Bethesda System criteria.[14] Because cases were chosen prospectively, no histologic follow-up was available at the time of the slide review.

Each of three pathologists (P.M.A., C.W.O., L.D.T.) reviewed all 100 cervical-vaginal smears twice, once by computer monitor and once by light microscopy. The delay between viewing of the digitized images and of the glass slides was approximately 2 weeks for each pathologist, and the observers

From the Division of Cytopathology, Department of Pathology, The Johns Hopkins Hospital, Baltimore, MD; and the Departments of Cellular Pathology, Otolaryngologic and Endocrine Pathology, and Telepathology, Armed Forces Institute of Pathology, Washington, DC.

The opinions or assertions herein represent the private views of the authors and are not to be construed as official or as representing the views of the Department of the Army, the Department of the Air Force, the Department of the Navy, or the Department of Defense.

Address correspondence and reprint requests to Timothy J. O'Leary, MD, PhD, Department of Cellular Pathology, Room G-137, Armed Forces Institute of Pathology, 1413 Research Blvd, Rockville, MD 20850.

0046-8177/01/3212-0005$0.00/0
doi:10.1053/hupa.2001.29651

did not consult each other regarding the study cases. Images were captured using a SONY DKC-5000 Digital Camera (SONY Medical Systems, Park Ridge, NJ) mounted on an Olympus BX40 microscope equipped with Olympus UplanFL objectives (Olympus America, Melville, NY). The camera was connected via SCSI Interface to a Micron 200 MHz Pentium CPU running Windows 98. One pathologist (I.K.) imaged the fields that had initially been "dotted" by the cytotechnologist screening the case. This pathologist did not participate in the subsequent dual review of the cases. The fields selected provided information routinely used in cytologic diagnosis, such as cellularity and background; reference cells, such as normal epithelium, neutrophils, and red cells; and representative fields showing abnormal features. Each area of interest was imaged at 200× and 400×. Images were captured at 1,544 × 1,120 resolution at 24-bit color depth and 300 dpi and imported directly into Adobe Photoshop 5.0 (Adobe Systems, Inc, San Jose, CA). Within Adobe Photoshop, image resolution was reduced to 72 dpi, and JPG compression of 10:1 was applied. Images from each slide were segregated into directories and transferred via file transfer protocol (FTP) to a Netscape Enterprise server (Netscape Communications Corp, Mountain View, CA). Images were viewed remotely over a local area network on workstations using Microsoft's Internet Explorer web browser (Fig 1). Conventional light microscopy was performed individually on an Olympus BH-2 microscope. Objective lenses were identical for each method.

Copies of the original cytology requisition and report forms were provided to each pathologist. The data supplied by the contributors varied by case, but all included patient age and variably included menstrual history, obstetric history, exogenous hormone status, and previous history of relevant cervical-vaginal lesions. In addition, the initial diagnosis rendered by the screening cytotechnologist was included on this requisition and report form. The final "sign-out" diagnosis was not included. The pathologists rendered diagnoses on each case using the Bethesda System.[14]

All pathologists were board certified in anatomic and clinical pathology by the American Board of Pathology, and 2 pathologists (observer 1 and observer 2) held additional qual-

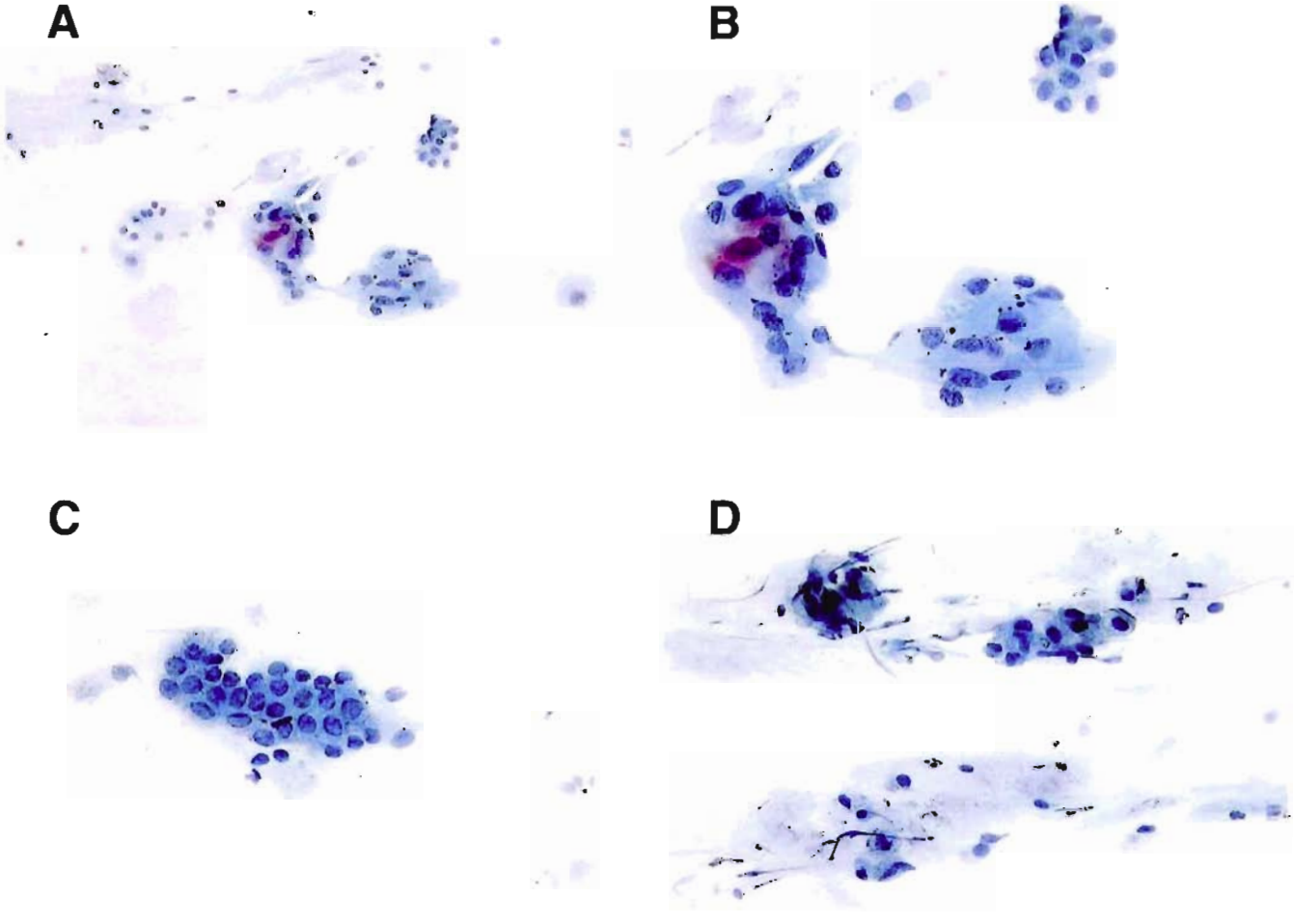

FIGURE 1. Typical images drawn from a single case in which 7 video fields were presented to the pathologists. Although eosin staining is slightly less apparent than in glass slides, cell borders remain apparent. Nuclear characteristics, including contours and chromatin structure, are readily appreciated. This cytotechnologist believed this case to be most appropriately classified as ASCUS. When viewing the glass slides, each of the pathologists agreed with this interpretation 2 favoring RCC, 1 favoring (LSIL). Two of the pathologists favored a diagnosis of LSIL based on the video images alone; the third preferred a diagnosis of ASCUS. (A) Image acquired with 20× objective lens. (B) Image of cell clusters from above, acquired with 40× objective lens. (C) Image acquired with 40× objective lens. (D) Image acquired with 40× objective lens. Same field as A with 40× objective lens.

ification in cytopathology from the American Board of Pathology. Observer 1 had significant experience, observer 2 moderate experience, and observer 3 little experience with telecytologic consultation.

To evaluate the diagnostic reproducibility between the telecytologic diagnosis and the glass slide diagnosis, the diagnoses were divided into 3 groups—benign (WNL/RCC), ASCUS/AGUS, and SIL (LSIL/HSIL)/carcinoma. These 3 categories were chosen because they reflect thresholds at which patient management decisions typically change significantly in the United States. Intraobserver and interobserver reproducibility were calculated using Cohen's κ statistic.[15] Intraobserver reproducibility between glass slide and video image diagnoses was calculated for each pathologist. Intersobserver reproducibilities for both glass slide diagnoses and video image diagnoses were calculated among all pathologists. Finally, interobserver reproducibility for both glass slide diagnoses and digital image diagnoses was calculated between each pathologist and the cytotechnologist. The cytologic diagnoses were considered semiquantitative[16] and were aggregated into 3 categories, with each diagnosis corresponding to a step from benign to malignant. The ordering of these 3 diagnostic categories was benign, ASCUS/AGUS, and SIL/carcinoma. Any difference in diagnostic category between glass slide and digitized images was considered a diagnostic disagreement. Therefore, telecytologic and light microscopic diagnoses that fell into the same category were considered concordant, and diagnoses that fell into different categories were considered discordant by 1 or 2 steps. Mean intraobserver reproducibility was calculated as the weighted average of the individual κ values.[17] To determine the significance of the interobserver κ values, we used the formula from Svanholm, et al.[15]

RESULTS

The intraobserver κ values between the digital images and glass slides for each of the 3 observers are shown in Table 1. A κ value of 1 reflects perfect agreement among all observers. When agreement is solely by chance, the κ value is 0, and at $\kappa < 0$ the observers generally disagree. Although there are no formal criteria by which to qualitatively describe κ values, many observers consider a $\kappa > 0.75$ to indicate excellent, κ of 0.58 to 0.74 good, κ of 0.4 to 0.57 fair, and κ of 0.2 to 0.39 poor reproducibility.[18] By these criteria, our results show fair to good intraobserver diagnostic reproducibility. The interobserver κ value for the 3 observers for the glass slides was 0.56 (variance, 0.00635). The interobserver κ value for the digital images was 0.45 (variance, 0.00641). The interobserver κ values between each of the 3 pathologists and the cytotechnologist's diagnosis are shown in Table 2. Although the diagnostic reproducibility was higher for glass slides (with κ values in the good to excellent range) than for digital images (κ values in the poor to good range), no 2-step diagnostic disagreements were found in this study.

TABLE 1. Intraobserver Reproducibility Between Digital Images and Glass Slides

Observer	Kappa (κ) Value
1	0.47
2	0.51
3	0.77
Mean	0.58

TABLE 2. Interobserver Reproducibility Between Observer and Cytotechnologist for Digital Images and Glass Slides

Observer	Observer/CT Glass Slide κ Value	Observer/CT Digital Image κ Value
1	0.59	0.32
2	0.77	0.46
3	0.67	0.58

Abbreviation: CT, cytotechnologist.

The disagreement between a pathologist's glass slide and digital diagnoses (mean intraobserver κ = 0.58) is less than that for different pathologists reviewing glass slides (slide interobserver κ = 0.56). Pathologists are even more likely to disagree on the interpretation of digital images (digital image interobserver κ = 0.45). Overall, diagnostic reproducibility was slightly higher for glass slides than for digital images.

DISCUSSION

Two different methodologic approaches have been advocated for telepathologic diagnosis. In dynamic systems, images are viewed live and in real time as the receiving viewer directly controls specimen orientation, field selection, and fine focus of the microscope via robotic controls.[18] In static systems, images are captured in a digital format on an image frame grabber board and then transmitted individually as still images to the receiving viewer. The receiving viewer usually has little or no direct control over microscope function.[10] Although dynamic imaging is unquestionably the more powerful technologic approach, the substantially lower cost favors the use of static imaging methods for review of cervical-vaginal smears (Fig 1).

When evaluating the reproducibility of diagnoses made via 2 or more different viewing modalities, the overall percentage or proportional agreement appears to be a simple and intuitively correct measure of reproducibility. Given the limited number of diagnostic possibilities, it is important to correct for chance agreement. *Agreement* is the overall or proportional number of cases given the same diagnosis between or within observers, including that part of the agreement, which may be attributable to chance. *Reproducibility*, that part of the agreement that may not be explained purely by chance, is appropriately measured by the κ statistic.[15] Reproducibility may be evaluated at the level of 2 or more observers examining the same specimen (interobserver reproducibility) or at the level of the same observer examining a specimen via 2 or more modalities or on 2 or more occasions (intraobserver reproducibility).

We found that the intraobserver and interobserver diagnostic reproducibility for both digital images and

the light microscope was fair to good, with κ values ranging from 0.47 to 0.77 (mean, 0.58) for intraobserver reproducibility and from 0.45 to 0.56 for interobserver reproducibility. The disagreement between a pathologist's glass slide and digital diagnoses (mean intraobserver $\kappa = 0.58$) is less than that for different pathologists reviewing glass slides (slide interobserver $\kappa = .56$). Pathologists are more likely to disagree on the interpretation of digital images (digital image interobserver $\kappa = 0.45$) than on the interpretation of glass slides. Overall, diagnostic reproducibility was slightly higher for glass slides than for digital images. It seems likely that the major factor underlying the better reproducibility of glass slide diagnoses among pathologists is the ability to review the entire pathologic specimen/slide before a diagnosis is rendered. Other factors, such as initial selection of slide fields for imaging and transmission, technical factors (digitization, transmission, and display), and viewer expertise and comfort with viewing and interpreting computer images would seem to play a greater role in determining intrapathologist disagreements in interpretation of glass slides and video images. As instrumentation improves and pathologists gain more experience in sending, receiving, and interpreting digital images, the diagnostic reproducibility of digital images will likely improve.

On the surface, our results differ somewhat from those of Raab et al[12] in their study examining the diagnostic accuracy of cytotechnologists.[12] These differences appear largely to reflect differences in the objectives and designs of our studies. Raab et al[12] reported κ statistics for the light microscope and video monitor of 0.34 and 0.20, respectively. However, their study compared the cytotechnologist's diagnosis on review of the video monitor or the light microscope with the original diagnosis rendered on each slide (interobserver variability). All slides had histologic confirmation of the cytologic diagnoses. In our study, the cytotechnologist/pathologist interobserver κ value ranged from 0.59 to 0.77 for slides and from 0.32 to 0.58 for digital images. In the study of Raab et al,[12] the diagnostic reproducibility was higher overall for glass slides than for digital images. In our study, all cytotechnologist diagnoses were based on glass slides. This will tend to lower cytotechnologist/pathologist interobserver κ values for the digital images. Additionally, the images in our study were digital, compared with the images viewed on the video monitor in the study of Raab *et al.* Finally, the groups of viewers being analyzed were different (cytotechnologists *v* pathologists). Thus, the apparent differences between the Raab et al study[12] and ours may reflect both differences in training and differences in diagnostic approaches to cervical-vaginal smears—screening versus rendering the final diagnosis.

In a study examining the diagnostic accuracy of video microscopy versus conventional examination of cervical-vaginal smears, Ziol et al[19] found an intraobserver κ of 0.47 to 0.81; these findings are similar to ours. However, in that study only 1 of the 6 participating pathologists reviewed all 100 study cases; the other 5 pathologists each reviewed 20 different cases. Thus it is difficult to interpet the resulting intraobserver κ values because values for each pathologist were calculated using different study cases. Compared with the reference diagnosis on each study case, Ziol et al report minimal difference in interobserver κ values for glass slides ($\kappa = 0.49$) and for video microscopy ($\kappa = 0.6$). However, in that study no interobserver κ value comparing the reviewing pathologists' results was determined.

The diagnostic categories used in this study are semiquantitative in nature; there were a limited number of categories, and they had an ordered nature to one another[16]: benign, ASCUS/AGUS, and SIL. For quality assurance pupose, a minor discrepancy is often defined as a 1-step difference between the original and observer diagnoses and a major discrepancy as a 2-step difference between the original and observer diagnoses. In the present study, no 2-step diagnostic disagreements occurred in any of the glass slide and/or digital image comparisons examined (pathologist intraobserver and interobserver κ and cytotechnologist/pathologist interobserver κ). This is a particularly encouraging finding because differences of more than 1 step may be expected to result in significantly different follow-up/treatment approaches. Ziol et al[19] report 4 2-step disagreements, reflecting the interpretation of very small cells that were difficult to visualize adequately using their video microscopy equipment.

A priori, there is reason to be reticent about rendering of diagnoses based on the review of only a few preselected fields as opposed to review and screening of the entire slide.[20] However, as pointed out by Raab et al,[12] the practice of rendering a cytologic diagnosis after reviewing abnormal fields dotted by an experienced cytotechnologist is well established. Our findings generally support the viability of remote pathologist review of selected abnormal fields of cytologic specimens in the context of a well-designed program in which women receive routine Pap smears. It seems likely to be less useful when women are screened infrequently. Our results have little or no bearing on questions relating to the use of video microscopy by cytotechnologists and in no way address issues related to use of video microscopy for quality control or proficiency testing purposes.

In summary, the intraobserver and interobserver diagnostic reproducibility of cervical-vaginal smears using digital images and the light microscope is fair to good. The use of remote digital images for pathologist review does not appear to pose a major risk of introducing large (2-step) diagnostic disagreements. These findings indicate that the use of telecytology for pathologist review of Pap smears holds much promise as a useful and reliable diagnostic tool.

REFERENCES

1. Weinstein RS, Bloom KJ, Rozek LS: Telepathology: Long-distance diagnosis. Am J Clin Pathol 91:S39-S42, 1989

2. Becker RL Jr, Specht CS, Jones R, et al: Use of remote video

microscopy (telepathology) as an adjunct to neurosurgical frozen section consultation. HUM PATHOL 24:909-911, 1993

3. Ito H, Adachi H, Taniyama K, et al: Telepathology is available for transplantation-pathology:experience in Japan using an integrated, low-cost, and high-quality system. Mod Pathol 7:801-805, 1994

4. Doolittle MH, Doolittle KW, Winkelman Z, et al: Color images in telepathology: How many colors do we need? HUM PATHOL 7:801-805, 1994

5. Kayser K: Telepathology in Europe. Its practical use. Arch Anat Cyto Pathol 43:196-199, 1995

6. Weinberg DS, Allaert FA, Dusserre P, et al: Telepathology diagnosis by means of digital still images: an international validation study. HUM PATHOL 27:111-118, 1996

7. Weinstien RS: Static image telepathology in perspective. HUM PATHOL 27:99-101, 1996

8. Weinstein, MH, Epstein JI:. Telepathology diagnosis of prostate needle biopsies. HUM PATHOL 28:22-29, 1997

9. Halliday BE, Bhattacharyya AK, Graham AR, et al: Diagnostic accuracy of an international static-imaging telepathology consultation service. HUM PATHOL 28:17-21, 1997

10. Weinstein LJ, Epstein JI, Edlow D, et al: Static image analysis of skin specimens: the application of telepathology to frozen section evaluation. HUM PATHOL 28:22-29, 1997

11. Weinstein RS, Bhattacharyya AK, Graham AR, et al: Telepathology: A ten year progress report. HUM PATHOL 28:1-7, 1997

12. Raab SS, Zaleski MS, Thomas PA, et al: Telecytology: Diagnostic accuracy in cervical-vaginal smears. Am J Clin Pathol 105:599-603, 1996

13. Briscoe D, Adair CF, Thompson LD, et al: Telecytologic diagnosis of breast fine needle aspiration biopsies. Acta Cytol 44:175-180, 2000

14. Kurman RJ, Solomon D: The Bethesda System for Reporting Cervical/Vaginal Cytologic Diagnoses. New York, NY, Springer-Verlag, 1994

15. Svanholm H, Starklint H, Gundersen HJG, et al: Reproducibility of histomorphologic diagnoses with special reference to the kappa statistic. APMIS 97:689-698, 1989

16. Raab SS, Bottles K, Cohen MB: Technology assessment in anatomic pathology: An illustration of technology assessment techniques in fine-needle aspiration biopsy. Arch Pathol Lab Med 18: 1173-1180, 1994

17. Fleiss JL: Statistical Methods for Rates and Proportions (ed 2). New York, NY, Wiley, 1981

18. Mun SK, Esayed AM, Tohme WG, et al: Teleradiology/telepathology requirements and implementation. J Med Sys 19:15-164, 1995

19. Ziol M, Vacher-Lavenu MC, Heudes D, et al: Expert consultation for cervical carcinoma smears. Reliability of selected field videomicroscopy. Anal Quant Cytol Histol 21:35-41, 1999

20. Mairinger T, Gschwendtner A: Telecytology using preselected fields of view: The future of cytodiagnosis or a dead end? Am J Clin Pathol 107:620-621, 1997